DATA+AI SUMMIT
BY databricks

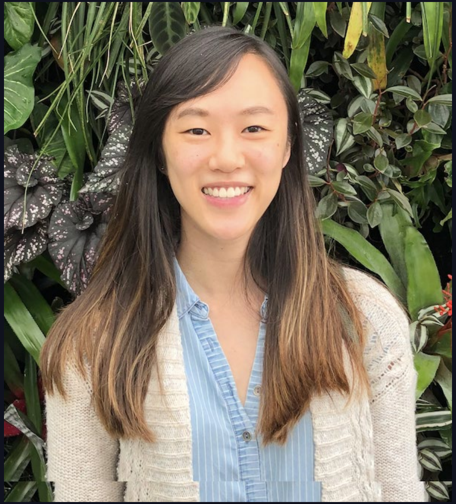
HOW CRISIS TEXT LINE'S DATA PLATFORM SUPPORTS MENTAL HEALTH

CRISIS TEXT LINE

Kristi Lui & Ankit Gupta

June 2024

Speakers:



Kristi Lui
Principal Product Manager

Ankit Gupta
Staff Data Engineer

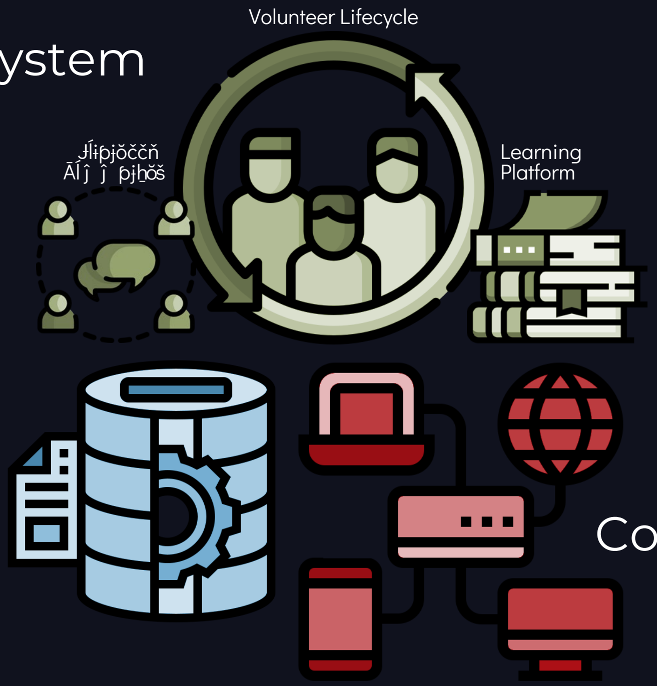
What is Crisis Text Line?

⚠

Crisis Text Line (and the technology that powers us) delivers unprecedented efficacy and efficiency to address the global mental health crisis.

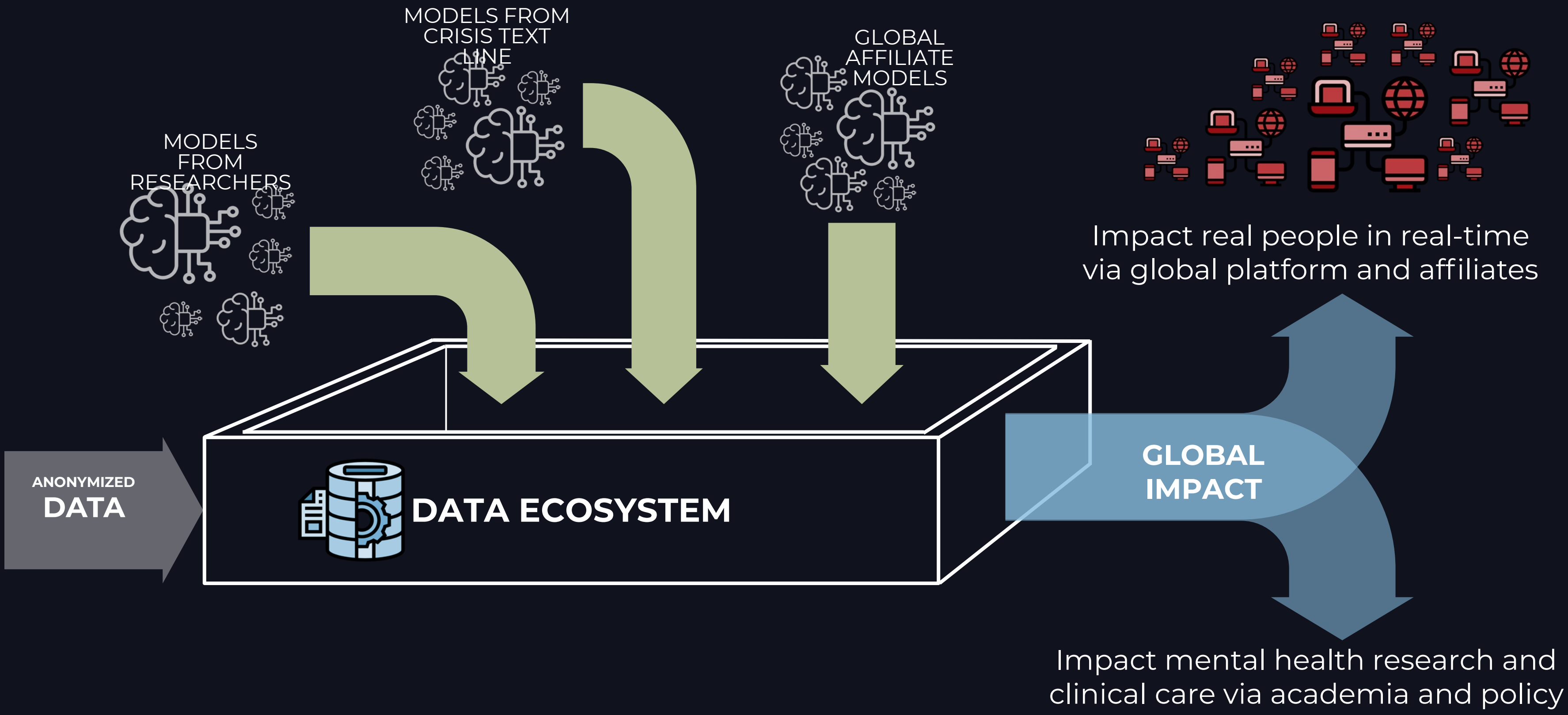
WHAT IS CRISIS CARE TECHNOLOGY?

Volunteer Ecosystem

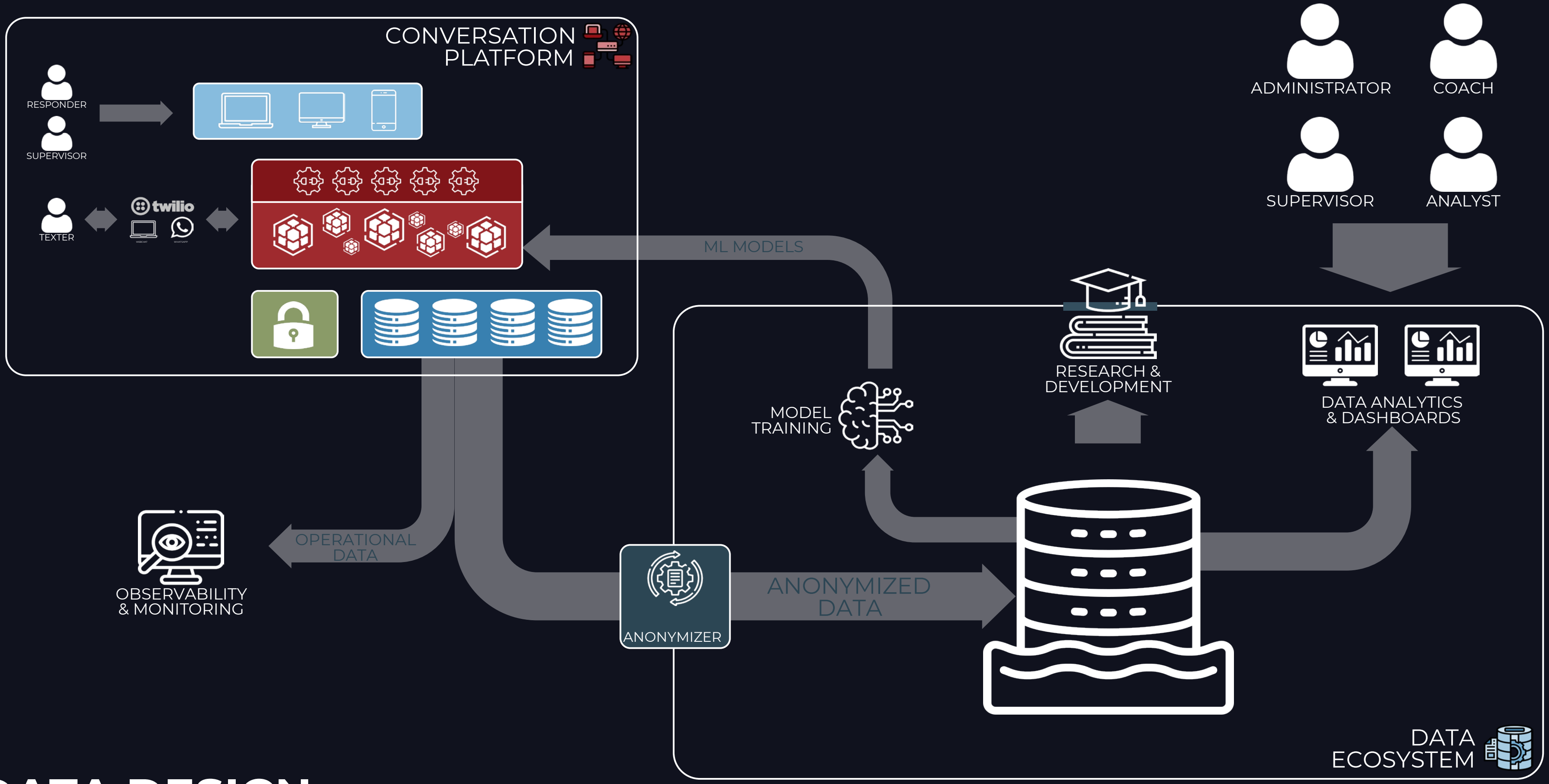


Data Ecosystem

Conversation Platform


MODELS FROM CRISIS TEXT LINE
GLOBAL AFFILIATE MODELS
MODELS FROM RESEARCHERS
Impact real people in real-time via global platform and affiliates
ANONYMIZED DATA
DATA ECOSYSTEM
GLOBAL IMPACT
Impact mental health research and clinical care via academia and policy

CRISIS CARE INNOVATION MARKETPLACE


CONVERSATION
PLATFORM
RESPONDER
SUPERVISOR
TEXTER
twilio
ML MODELS
OPERATIONAL
DATA
OBSERVABILITY
& MONITORING
ANONYMIZER
ANONYMIZED
DATA
MODEL
TRAINING
RESEARCH &
DEVELOPMENT
DATA ANALYTICS
& DASHBOARDS
ADMINISTRATOR
COACH
SUPERVISOR
ANALYST
DATA
ECOSYSTEM

DATA DESIGN

This new data platform focuses on centralizing and securing:

- Texter, conversation, & survey data
- Responder and staff data
- Content & Resource data
- Operational data, including monitoring and logs
- Data Pipelines to access the data

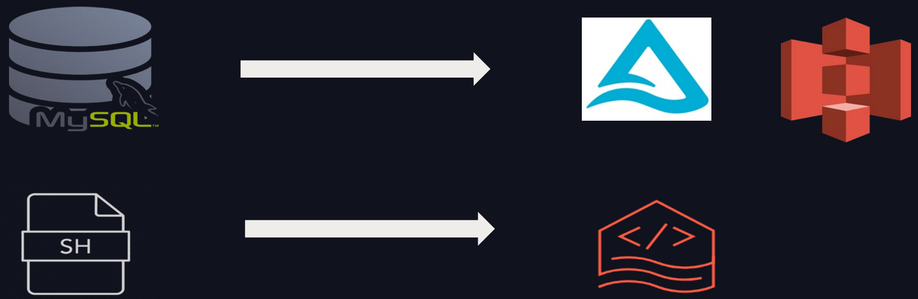
How do we accomplish this?

- New and modern technology stack
- Federated datastores without synchronization and duplication
- Dedicated data stack
- Clear separation between operational and analytics data

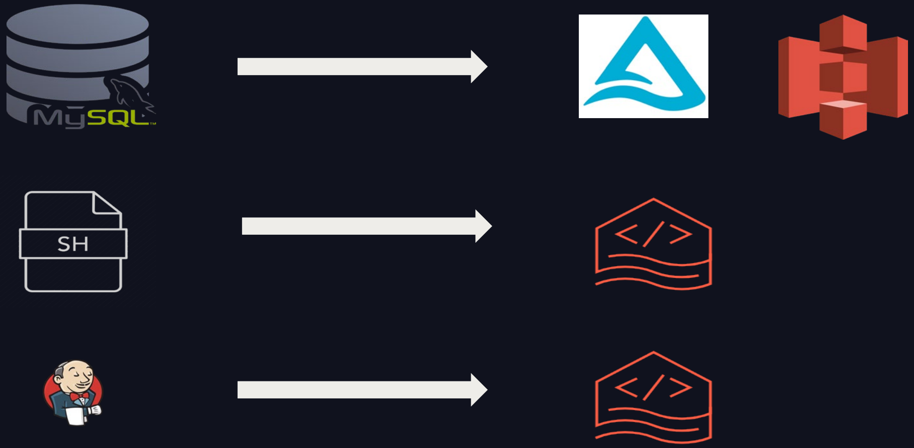
The Evolution

Data Storage

Data Pipeline



The Evolution



Data Storage

Data Pipeline

Data Orchestration

The Evolution

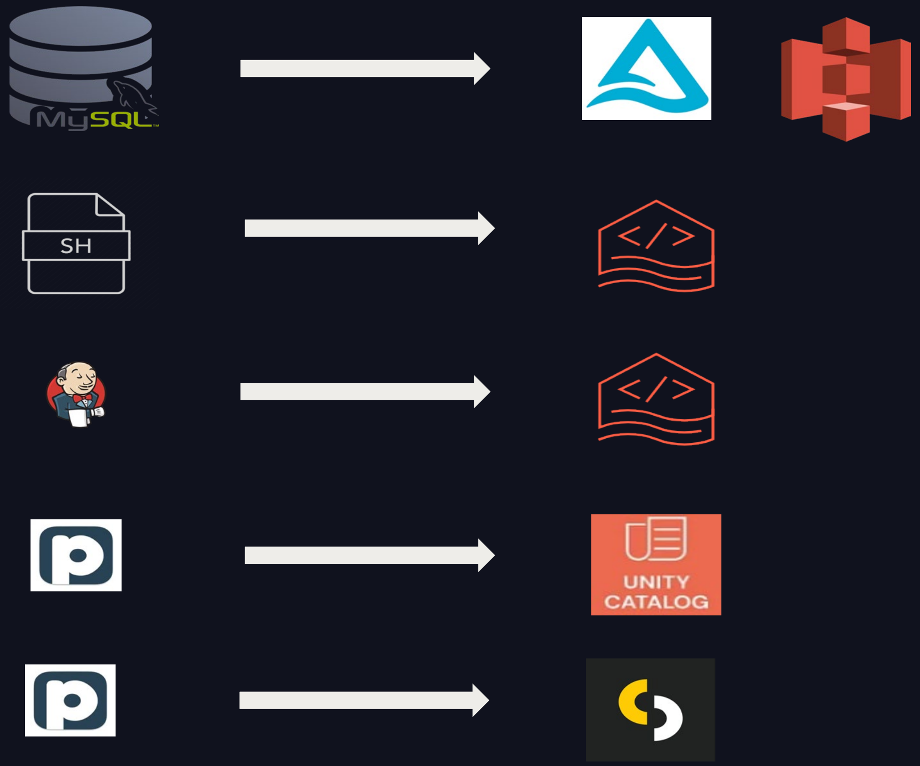
Data Storage

Data Pipeline

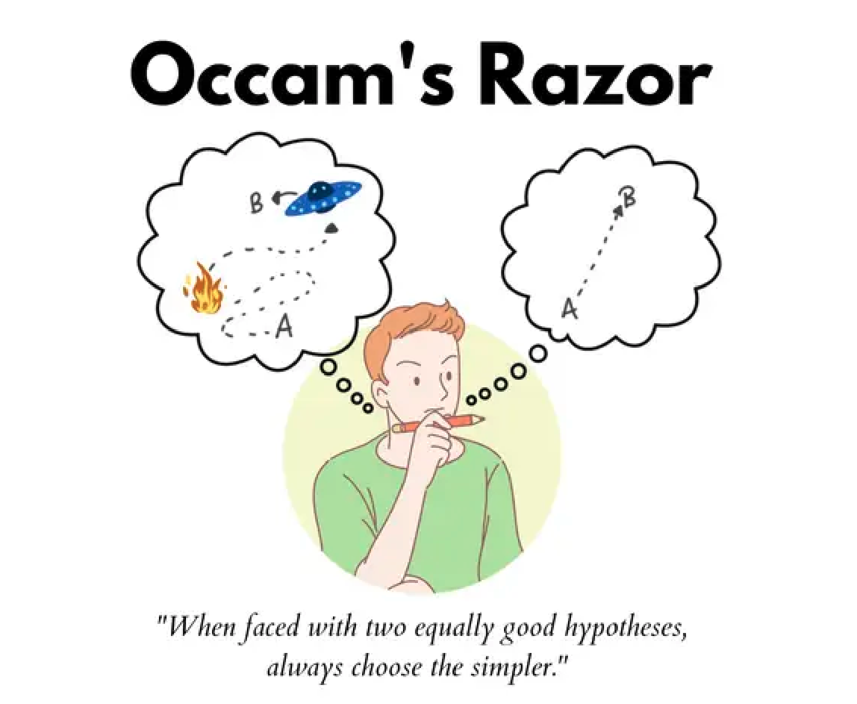
Data Orchestration

Data Access

Business Intelligence

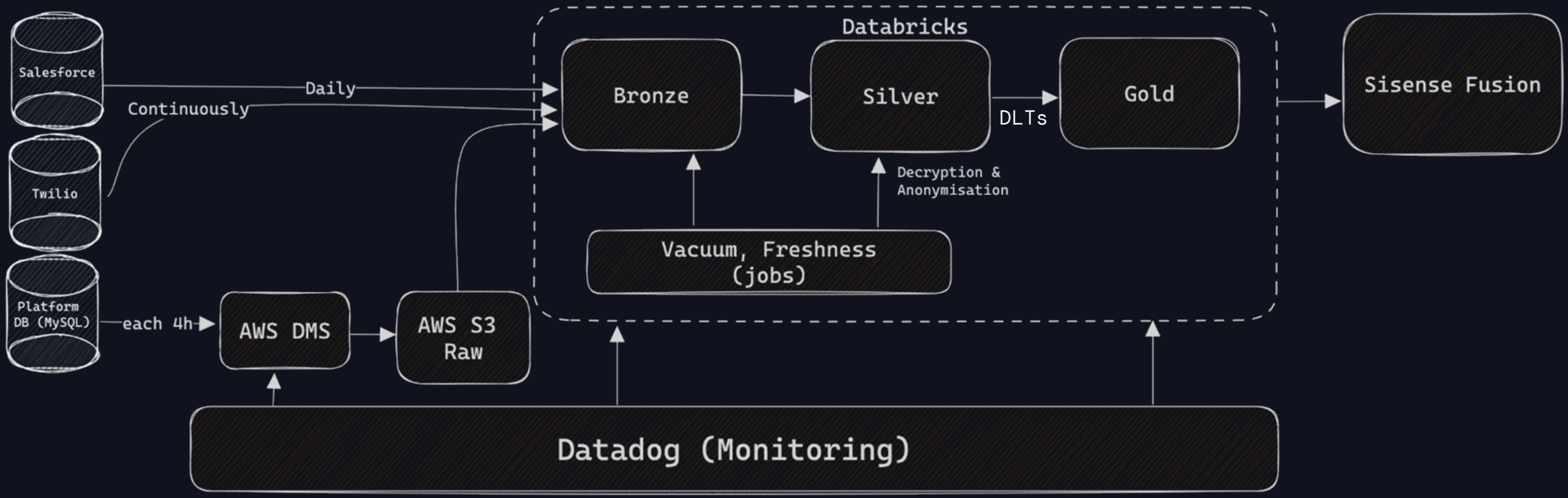


10% steps to 10x goals


Occam's Razor
B
A
B
A
"When faced with two equally good hypotheses, always choose the simpler."

source

The goal



Monitors:
- Tables processed & failures
- Data freshness

Bronze and Silver Layers

SPARK JOBS. Guiding principles

1. Created PySpark applications for transformation and ingestion
2. Pure functions for transformation logic, which allows us to simplify code, e.g. `df.transform(func_1).transform(func_1)`
3. Handles Unit Tests, part of the CI/CD pipeline.
4. Job parameters as yaml-files for each workspace.
5. Data asset definitions yaml-files, soon to be data contracts.
6. Notebooks are used to start Spark applications triggered by Databricks workflow and by Analytic teams with full responsibility.

Bronze and Silver Layers

SPARK JOBS. Sample Python Code

PYTHON (notebook-wrapper)

```
# Databricks notebook source
import sys

from utils.input_parameters import create_widget_for_env,
get_parameters_for_runner

create_widget_for_env(dbutils)

# COMMAND ----------
from spark.raw_to_landing.twilio.twilio_ingest import TwilioIngest

try:a
   app_name = dbutils.widgets.get("app_name")
except Exception as e:
   e = str(e)
   if 'InputWidgetNotDefined' in e:
       app_name = "twilio_phoenix_ingest_job"

job_params = get_parameters_for_runner(dbutils, app_name)

config =
job_params.config_yaml["stages"]["raw_to_landing"]["spark_jobs"]["twilio_phoeni
x_ingest_job"]
TwilioIngest().main(**config, spark=job_params.spark)
```

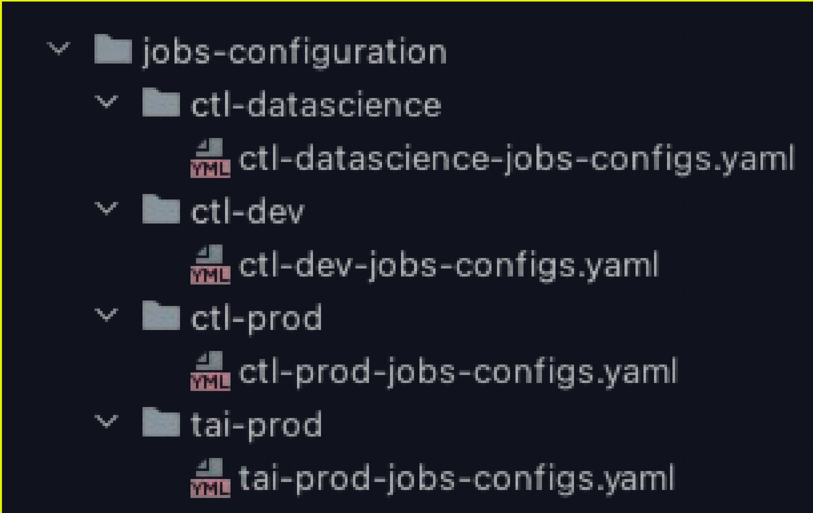
PYTHON (main-job)

```
DataFrameTransform = Callable[[DataFrame], DataFrame]

class TwilioIngest(SparkCommonTransformations, TwilioHelper):

   def main(self,
            job_input: Dict,
            job_output: Dict,
            operational_monitoring: Dict,
            app_name: str,
            spark: Optional[SparkSession] = None,
            **kwargs):
             # implementation …
       result_df = (df
                    .transform(self.with_current_ts)
                    .transform(with_md5_hash(md5_columns))
                   )
```

Spark Jobs

Jobs config

STRUCTURE



```
jobs-configuration
  ctl-datascience
    ctl-datascience-jobs-configs.yaml
  ctl-dev
    ctl-dev-jobs-configs.yaml
  ctl-prod
    ctl-prod-jobs-configs.yaml
  tai-prod
    tai-prod-jobs-configs.yaml
```

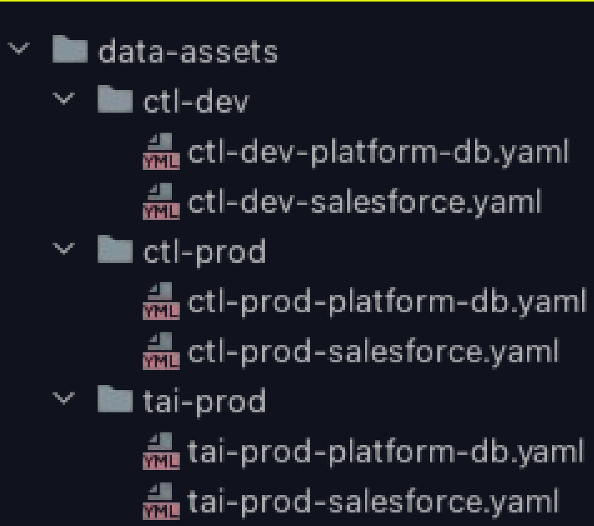
CONFIG-yaml

```
stages:
 raw_to_landing:
   spark_jobs:
     full_load_job:
       app_name: full_load_job
       job_input:
         datadog_api_key_config:
           <<: *datadog_api_key
         load_base_path: "s3://XXXXXXX"
           tables_for_processing: *platform_db_tables
         data_assets_config: *data_assets_config
       job_output:
         schema: ctl_dev_bronze.platformdb
       operational_monitoring:
         labels:
           <<: *OM_global_labels
           job_name: full_load_job
```

Spark Jobs

Data-assets-yaml

STRUCTURE

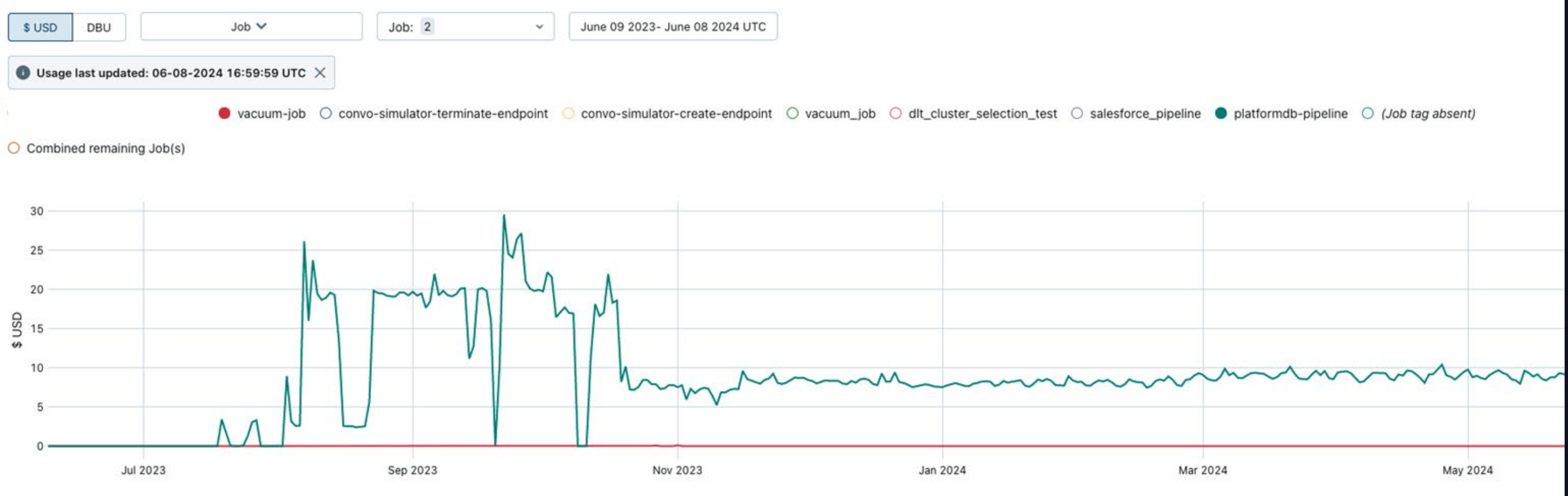


```
data-assets
  ctl-dev
    ctl-dev-platform-db.yaml
    ctl-dev-salesforce.yaml
  ctl-prod
    ctl-prod-platform-db.yaml
    ctl-prod-salesforce.yaml
  tai-prod
    tai-prod-platform-db.yaml
    tai-prod-salesforce.yaml
```

CONFIG-yaml

```
message:
 identity:
   - id
 encrypted:
   - message
 scrubbing:
   - message
 md5_checksum: [message, status, type, timestamp]
reporting_log_value:
 identity:
   - id
 scrubbing:
   - value
 md5_checksum: [question_id, value, group_number]
```

Solve then scale



$ USD | DBU

Job

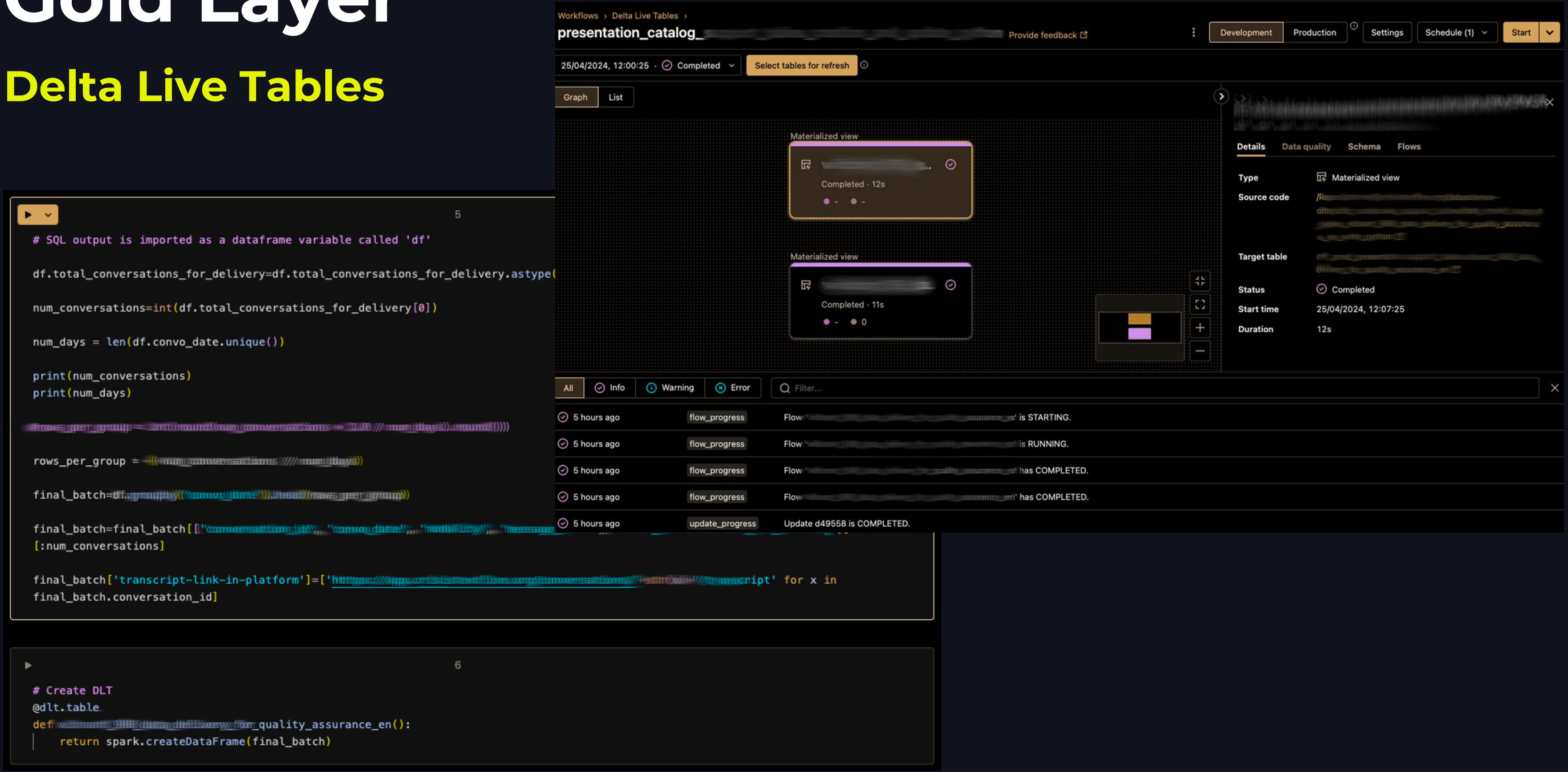
Job: 2

June 09 2023- June 08 2024 UTC

Usage last updated: 06-08-2024 16:59:59 UTC

Gold Layer

Delta Live Tables



```
# SQL output is imported as a dataframe variable called 'df'

df.total_conversations_for_delivery=df.total_conversations_for_delivery.astype(

num_conversations=int(df.total_conversations_for_delivery[0])

num_days = len(df.convo_date.unique())

print(num_conversations)
print(num_days)

[redacted]

rows_per_group = [redacted]

final_batch=[redacted]

final_batch=final_batch[[redacted]
[:num_conversations]

final_batch['transcript-link-in-platform']=['[redacted]script' for x in
final_batch.conversation_id]
```

```
# Create DLT
@dlt.table
def [redacted]_quality_assurance_en():
    return spark.createDataFrame(final_batch)
```

Workflows › Delta Live Tables ›

presentation_catalog_[redacted] Provide feedback

Development | Production | Settings | Schedule (1) | Start

25/04/2024, 12:00:25 · Completed | Select tables for refresh

Graph | List

Materialized view
[redacted]
Completed · 12s

Materialized view
[redacted]
Completed · 11s

[redacted]

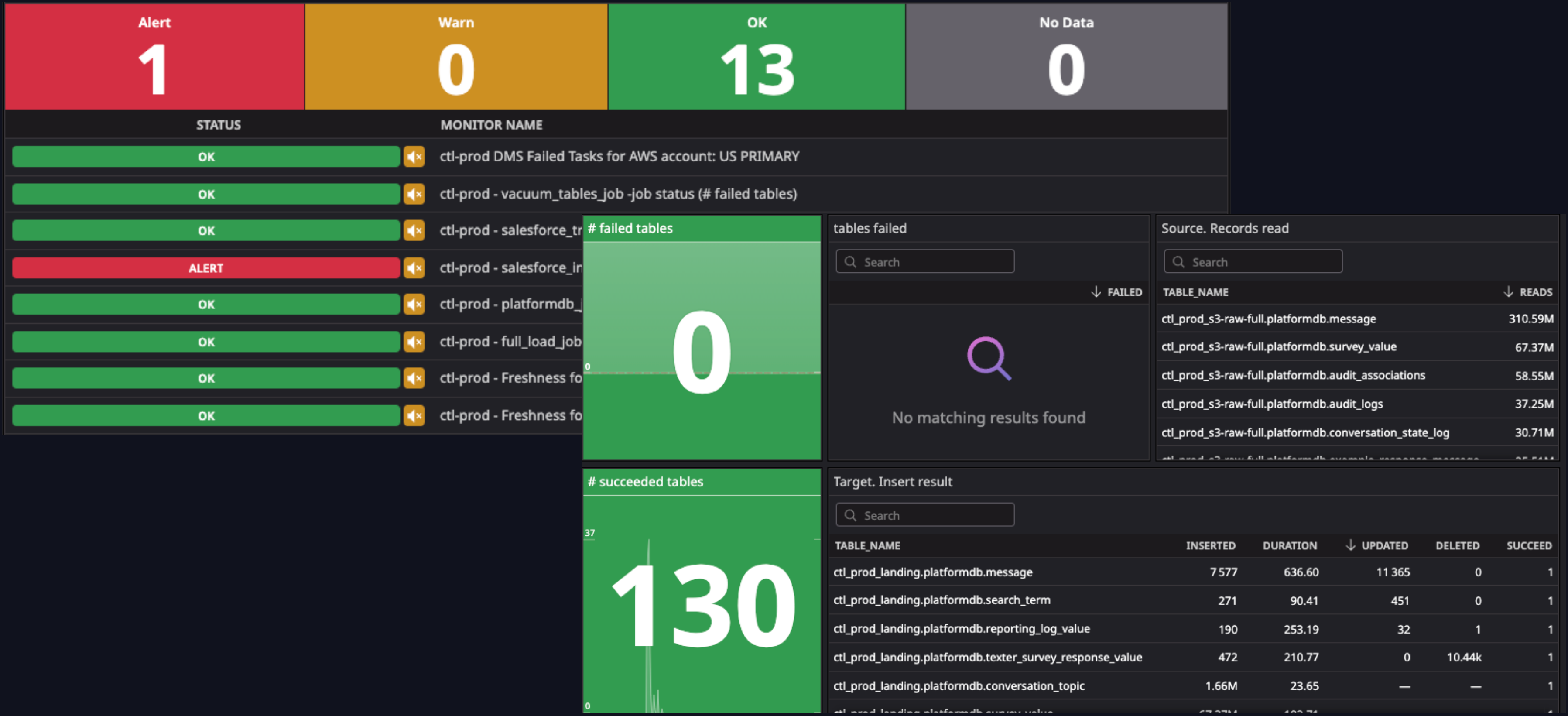
Details | Data quality | Schema | Flows

Type	Materialized view
Source code	[redacted]
Target table	[redacted]
Status	Completed
Start time	25/04/2024, 12:07:25
Duration	12s

All | Info | Warning | Error | Filter...

```
5 hours ago    flow_progress      Flow '[redacted]' is STARTING.
5 hours ago    flow_progress      Flow '[redacted] is RUNNING.
5 hours ago    flow_progress      Flow '[redacted]' has COMPLETED.
5 hours ago    flow_progress      Flow '[redacted]' has COMPLETED.
5 hours ago    update_progress    Update d49558 is COMPLETED.
```

System Health



Alert	Warn	OK	No Data
1	0	13	0

STATUS	MONITOR NAME
OK	ctl-prod DMS Failed Tasks for AWS account: US PRIMARY
OK	ctl-prod - vacuum_tables_job -job status (# failed tables)
OK	ctl-prod - salesforce_tr
ALERT	ctl-prod - salesforce_in
OK	ctl-prod - platformdb_j
OK	ctl-prod - full_load_job
OK	ctl-prod - Freshness fo
OK	ctl-prod - Freshness fo

failed tables

0

tables failed

Search

↓ FAILED

No matching results found

Source. Records read

Search

TABLE_NAME	↓ READS
ctl_prod_s3-raw-full.platformdb.message	310.59M
ctl_prod_s3-raw-full.platformdb.survey_value	67.37M
ctl_prod_s3-raw-full.platformdb.audit_associations	58.55M
ctl_prod_s3-raw-full.platformdb.audit_logs	37.25M
ctl_prod_s3-raw-full.platformdb.conversation_state_log	30.71M

succeeded tables

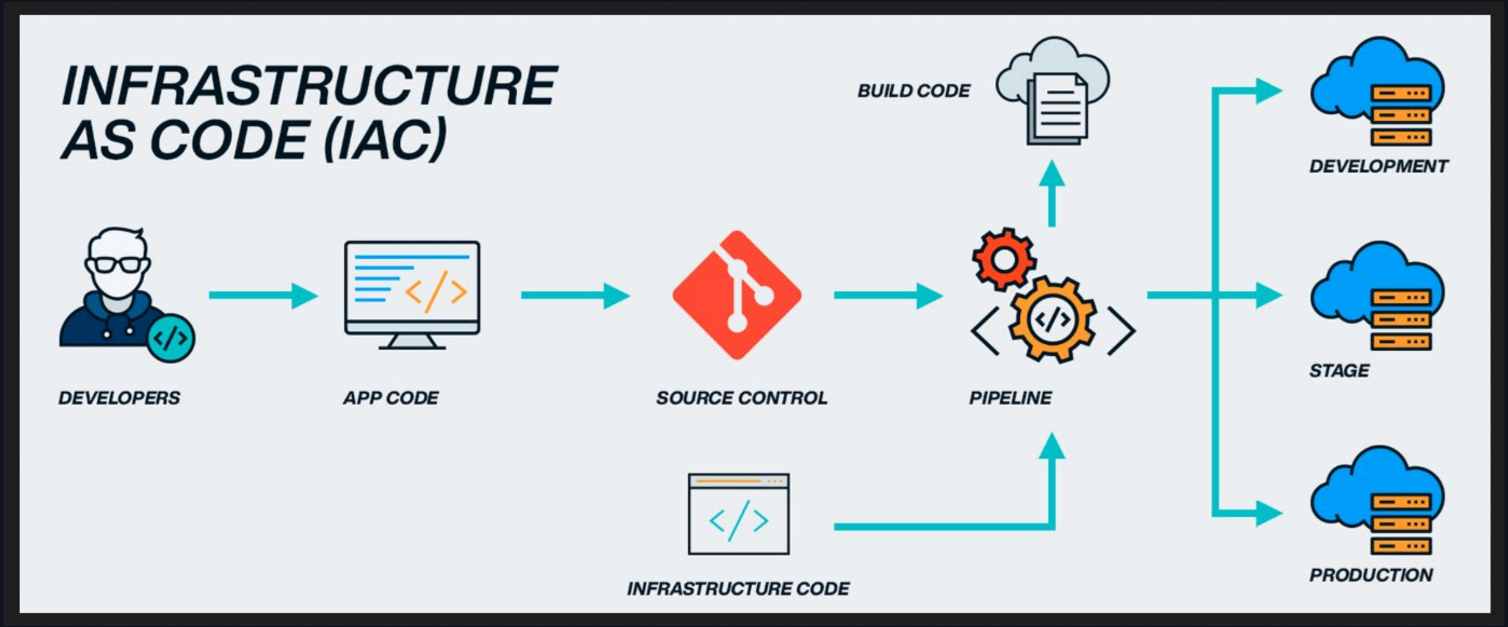
130

Target. Insert result

Search

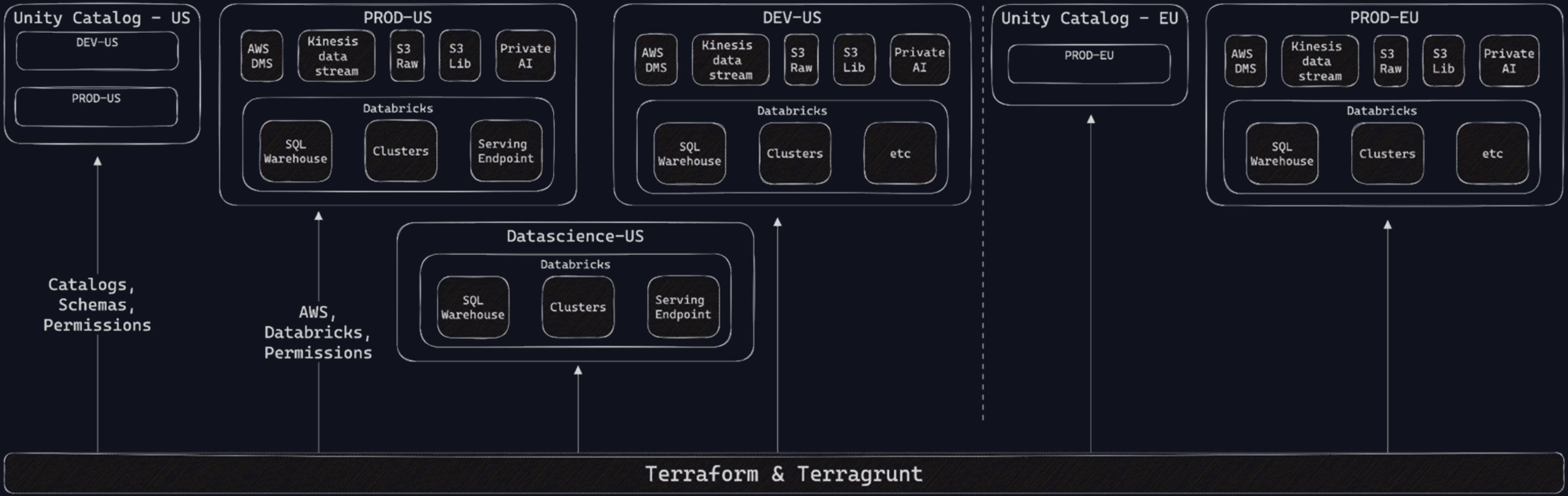
TABLE_NAME	INSERTED	DURATION	↓ UPDATED	DELETED	SUCCEED
ctl_prod_landing.platformdb.message	7 577	636.60	11 365	0	1
ctl_prod_landing.platformdb.search_term	271	90.41	451	0	1
ctl_prod_landing.platformdb.reporting_log_value	190	253.19	32	1	1
ctl_prod_landing.platformdb.texter_survey_response_value	472	210.77	0	10.44k	1
ctl_prod_landing.platformdb.conversation_topic	1.66M	23.65	—	—	1

Deployment at scale



source

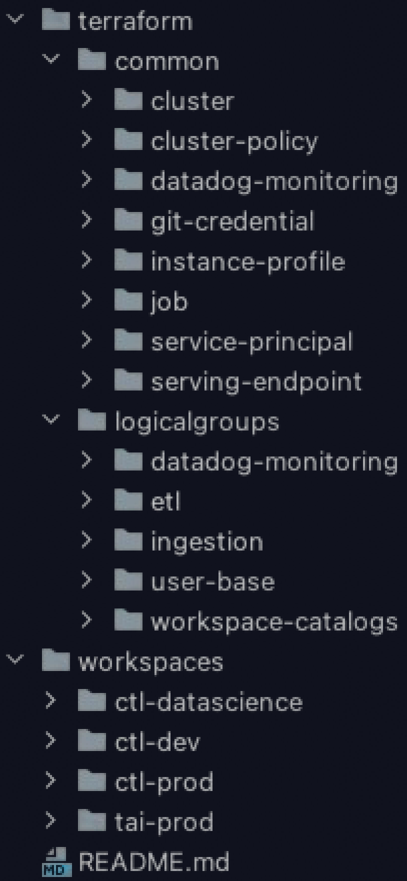
Deployment at scale



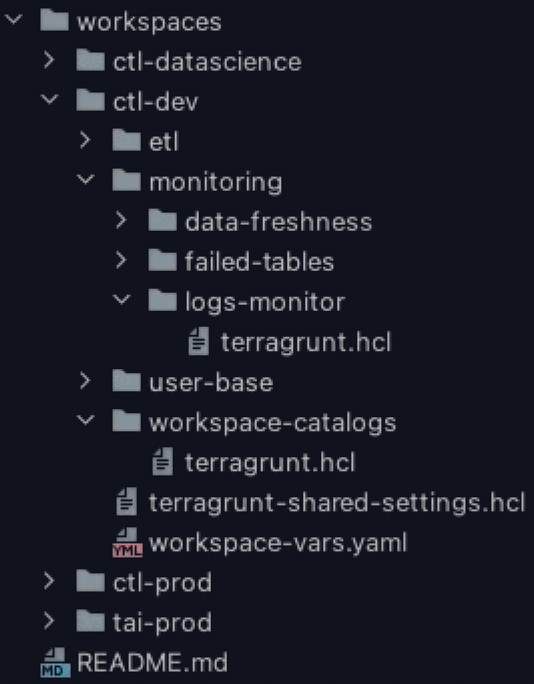
Deployment at scale

Terraform covers

- Clusters, Cluster-policies
- Datadog
- ETL jobs
- Service principals
- Catalogs, schemas
- Permissions

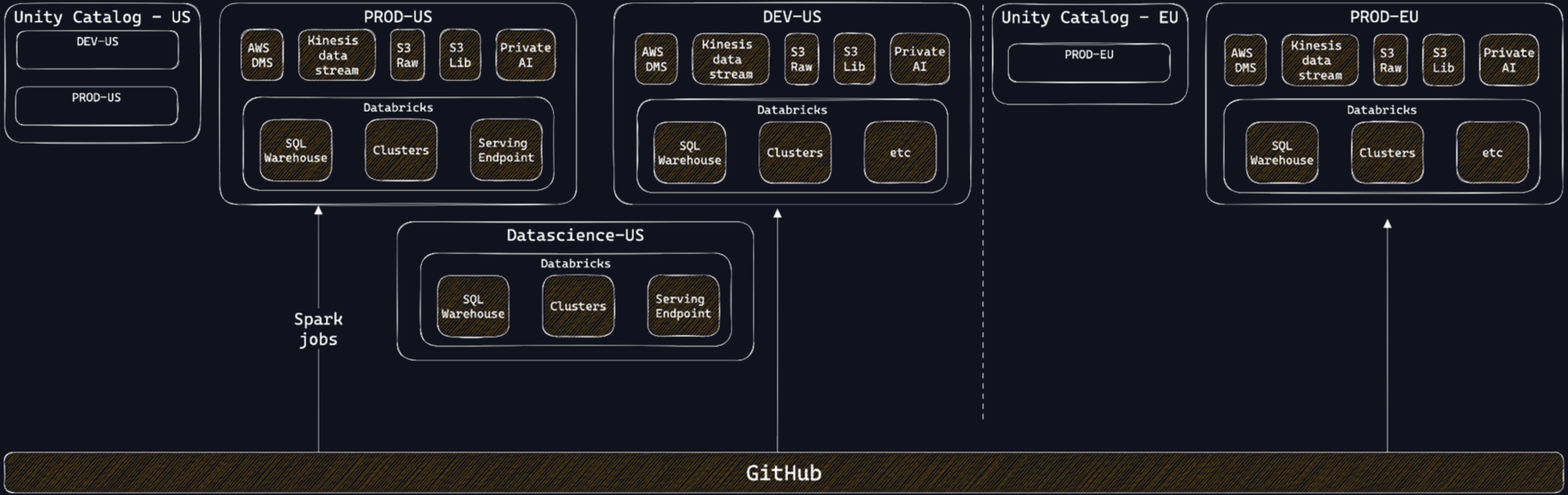


```
terraform
  common
    cluster
    cluster-policy
    datadog-monitoring
    git-credential
    instance-profile
    job
    service-principal
    serving-endpoint
  logicalgroups
    datadog-monitoring
    etl
    ingestion
    user-base
    workspace-catalogs
workspaces
  ctl-datascience
  ctl-dev
  ctl-prod
  tai-prod
README.md
```

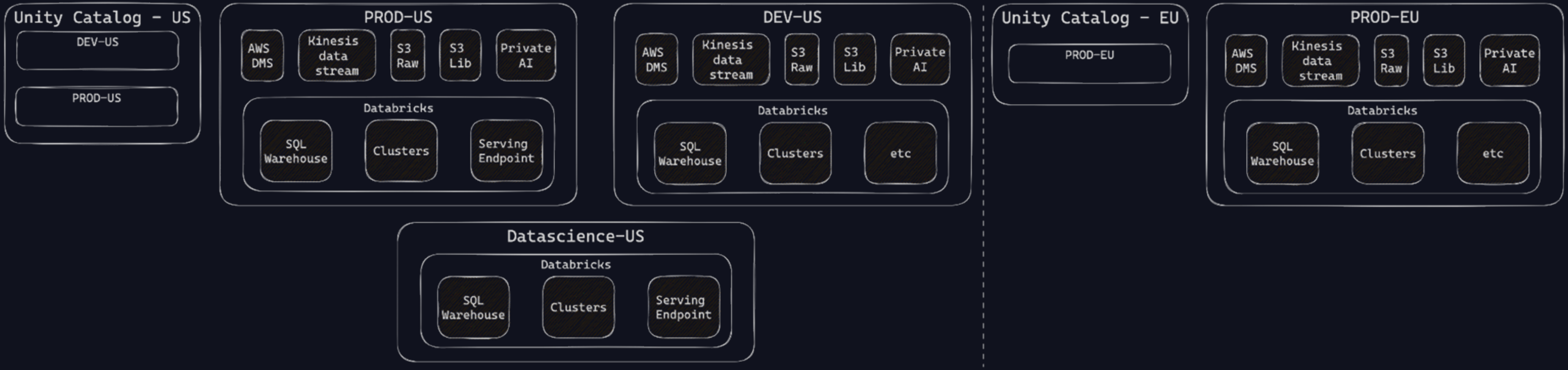


```
workspaces
  ctl-datascience
  ctl-dev
    etl
    monitoring
      data-freshness
      failed-tables
      logs-monitor
        terragrunt.hcl
    user-base
    workspace-catalogs
      terragrunt.hcl
    terragrunt-shared-settings.hcl
    workspace-vars.yaml
  ctl-prod
  tai-prod
README.md
```

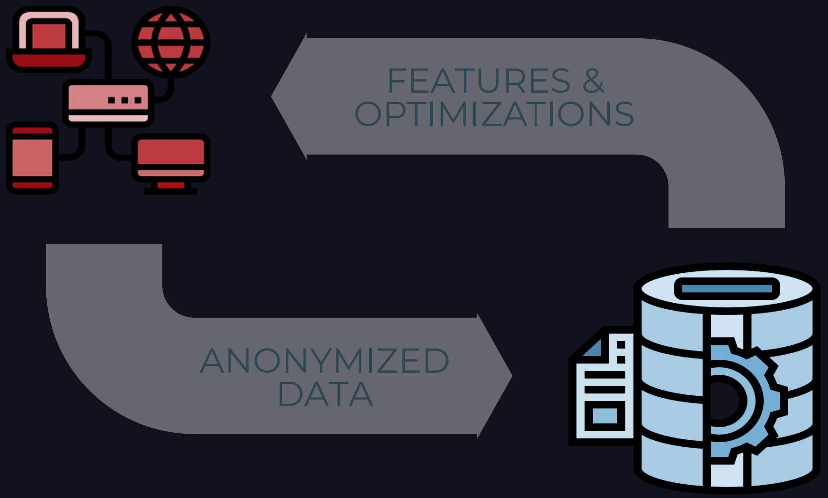
Deployment at scale



Scaling beyond US



HOW TO BUILD THE FUTURE OF CRISIS SUPPORT



A LEARNING SYSTEM

Our data scientists and researchers actively accelerate the journey of crisis support by using our data to improve our platform

Our Data Platform, with 10 years of anonymized data, powers the models and science that optimizes and accelerates our service

DATA, ML, AI

OPERATIONAL EFFICIENCY

Ensure responders and supervisors are meeting demand and maximizing impact

RISK DETECTION & TRIAGE

Analyze conversations to flag active risks and prioritize support

RESPONDER AUGMENTATION

Realtime support for responders to streamline conversations and improve quality

CLINICAL ANALYTICS

Evaluate care performance to ensure high quality support. Responsiveness to changing patterns.

GLOBAL IMPACT

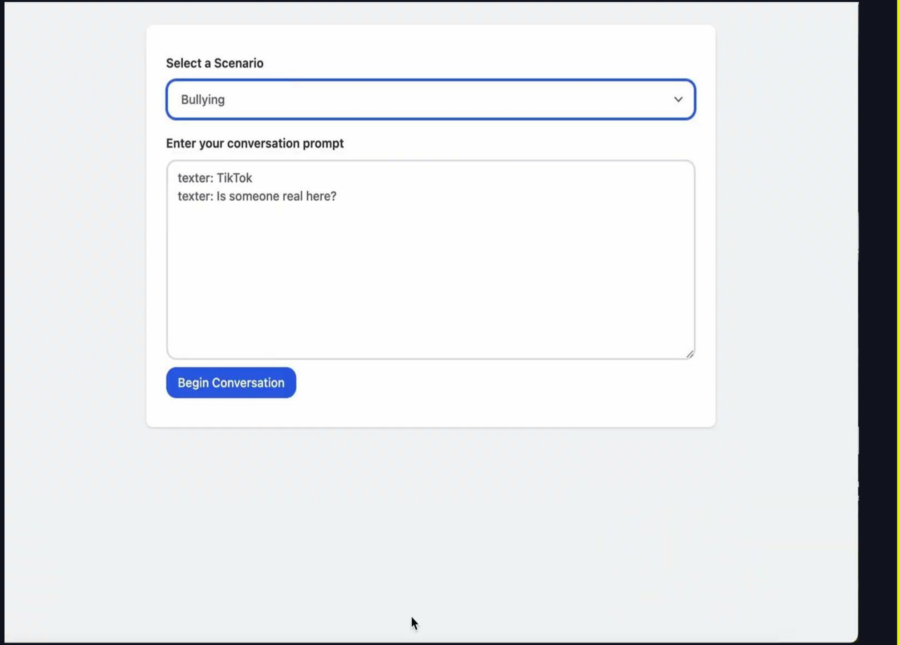
Draw upon worldwide crisis care data, and project learning and models across the planet

TRAINING & EVOLUTION

Ensure high care quality via data-driven simulation and augmentation

GenAI for Mental Health

Conversation Simulator



Select a Scenario

Bullying

Enter your conversation prompt

texter: TikTok
texter: Is someone real here?

The world needs more human to human interaction, not less.

- Experience
- Low risk
- Continuous learning

Thank you

Sign up to be a volunteer!

Learn more about us!

Q&A